The Trouble with Numbers
(Interpreting Data)

"We decided to total-out your wife's car" the insurance agent reported, (not surprising as I recalled the image of the streetlight looming out of the dashboard of the '92 Caprice) "…and we have assessed the car's value at $4000." Now that one stopped me. The NADA showed $5000 and my brief newspaper search had resulted in only a '91 and a '93, both priced well above $5000. After voicing my concerns to the agent he replied that they didn't use NADA, but rather had a company that provided them with prices on similar cars. I suggested that his listings might represent higher than average mileage (which at the time was about 89,000 miles), but he assured me that at worst case the cars they priced had average mileage. At my request he provided me with phone numbers for the cars they used, and surprisingly there were only three listings. Even more unusual I thought was that two of the cars were located 200 miles away from my local area. However, the most intriguing part was the cars themselves. One car had 131,000 miles. Another car was a police cruiser with 300,000 miles! And yes, this is a true story.

If the above account were an isolated incident there would be no cause for concern. Unfortunately misuses of data occur countless times each day. In fact, the era we live in could legitimately be called the "Misinformation Age". A recent article in Consumers' Research[1] stated that while polls and surveys have greatly increased in number, the accuracy of such studies has actually declined. The media often sacrifices precision for speed in their eagerness to make statements on public opinion, report trends in the market place, or draw attention to social and environmental issues. Errors are further compounded when such studies are taken out of context and are used to make broad generalizations.

Never has the warning "caveat emptor", let the buyer beware, been more fitting, and how much more so for those people charged with the responsibility for cost and price analysis. These are the people known as business managers, financial managers, contracting officers, buyers, pricers, estimators and various other titles that I will simply refer to as analysts. Adding to the challenges of today's analysts is the growing pressure to make faster decisions with less data and yet with more confidence. The only way to gain this confidence is through a better understanding of the numbers with which we work and the factors that effect them.

Sampling – A Chip Off The Old Block

The use of samples to make statements or inferences about the populations from which they were drawn is a common practice. The Gallup Organization[2] typically uses a sample size of 1,000 adults to make inferences about the opinions of a population of 187,000,000 adults, and generally with an accuracy of ± 3 percent. So where did the insurance agent in the opening story go wrong?

Sampling is one area where size really does matter. The larger the sample the more representative it is of the population. Larger sample sizes allow the analyst to employ a variety of graphical and computational diagnostic techniques to better understand the nature of the data.

---

[1] Consumers' Research, A Consumer Guide To Opinion Polls, September 2000.
[2] The Gallup Organization, How Polls Are Conducted, FAQ.

Conversely, the small sample of three cars in my example makes it very difficult to state with any certainty the properties of the population of '92 Caprices. Now, while this sounds like we should collect as much data as possible, judgment is required in balancing between the additional confidence and accuracy that comes with larger sample sizes and with the resources required to obtain the additional data. For example, let's say that you are purchasing toner cartridges for your office printers. Checking prices with three vendors would probably be sufficient given that a toner cartridge is very well defined and given the small dollar amount associated with the purchase. Next, let's consider that you are contracting for roof repair for your organization. How large is the roof? Are you repairing one building or a number of buildings? Is there some question as to whether the roof should be patched or replaced? If the roofing needs to be replaced, do you replace the roofing itself or does structural work need to be accomplished as well? Is there a fair amount of variation in the price and quality of the materials? Should a life cycle cost and benefit analysis be done? This situation calls for more data than the toner cartridges because of the dollar value involved and the uncertainty associated with the actual requirement. It would also call for more market research and data analysis to include consulting with civil engineering, an independent consultant, etc. Again, it's a balancing act between confidence and resources. Remember the old adage, "There's strength in numbers".

Another important principle in sampling is that the sample should be random in nature, meaning that any observation in the population should have an equal opportunity of being selected. Some market research on my part revealed that the insurance company's sample car prices were among the lowest in the market place, a sample that was anything but random. Although this may sound obvious, a common oversight is not checking your sample to make sure all of the items came from the population in question. I think we can conclude that in the case of my insurance claim that a police cruiser did not belong in a sample of personal vehicles. This may not always be so straightforward. Do HF, UHF, and VHF all belong to the same population of radios? Are there differences between HF radios installed in airplanes, helicopters, and tanks? The best advice in constructing or evaluating samples is to consult with someone who has experience in that particular area. And while it is important to ask why each item was included in the sample, it is just as important to ask why certain items were excluded. For example, in a recent aircraft acquisition the contractor provided the government with the prices of similar aircraft to support the fairness of their price. What was absent, however, was the price of the previous version of the aircraft being procured. A suspicious person might speculate that the price of the previous version was omitted to avoid drawing attention to the significant disparity in the prices of the two versions.

Why does my apple juice taste like fruit punch?

It would be nice to think that having selected an appropriate sample we would be ready to start crunching numbers, but we need to first stop and ask ourselves if we are making "apples to apples" comparisons. Numbers, unfortunately, do not come with a list of ingredients. We need to determine what the numbers represent. Your car gets 30 miles per gallon. Is that highway, city, or a combination of the two? The price of the car is $17,000. Does that include warranty, taxes, and tags? Careful attention to the content that the numbers represent is essential to avoid getting fruit punch when you were trying to make apple juice.

Some of the problems that Fisher[3] identified when it comes to comparing numbers deal with definitions, formats, matching-up, and temporal factors. Definitional problems occur when terms are subjective in nature or when terms can be used in various contexts. What are nonrecurring costs, and should they be considered a development cost, a cost spread between development and the first lot of production, or should the nonrecurring costs be amortized over the production run? You might think that there would be little confusion in referring to the weight of an aircraft, however, as it turns out there are different kinds of weight such as airframe unit weight, empty weight, gross take-off weight, etc. Using a number should be like following a recipe, pay close attention to the ingredients and how they should be used.

Have you ever tried to price cars and found some were advertised as $300 a month, some as $14,500, and yet others as $12,500 with trade-in? Did you want to know the warranty cost, but were told that it was "not separately priced". The problem in this case is that the information is not in a format suitable for making comparisons. Format problems can occur when information is in different types of formats or when information is in different levels of format. You may want cost reporting on an item, part by part, but the contractor records cost by manufacturing and engineering categories. It may be difficult to compare lower level costs like data, spares, logistics support, and warranties across various contractors because their accounting systems may allocate the costs differently. You may want to know the software development cost for an electronic component, but find that the software costs were not segregated from the hardware costs. If you were buying jet engines you might be interested in the price of turbines, but find that the cost is not broken down to that level of detail. Maybe you've collected prices on some product or service, but find that the prices represent different quantities or levels of effort or periods of time. In each case a decision must be made as to whether the data can be adjusted to account for the differences and what level of error, if any, is introduced as part of making the adjustment. Depending on the level of confidence you are looking for, you may choose not to use certain data points because you cannot accurately adjust the data into the desired format.

On the "Price is Right" contestants are often called upon to match prices to products, a "matching" problem we often have to face ourselves. Let's say that you wanted to develop a cost estimating relationship between an airframe cost and the airframe unit weight. You find out that the F-16 airframe cost $14,000,000 and the weight is 12,000 pounds. While this sounds simple enough, there is a problem. Both the cost and the weight are moving targets. Is the cost associated with the first aircraft built, the 100th, 200th, or the 500th aircraft built? Was there a learning curve where the cost decreased in real dollars as more units were produced due to improvements during production? Does the cost represent an average cost for a given production lot? Is the average cost affected by variations in quantity? The weight is no more static in nature than the cost because over the production life of an aircraft it typically experiences weight growth. Given the tendency for these values to vary, do we even know if the cost and weight reflect the appropriate values for the same aircraft? Data analysis is like a Rubik's Cube, you're not done until everything matches.

Temporal problems occur because things change over time. One effect on prices over time is inflation. Care must be taken to avoid comparing prices from different years without first having adjusted the prices to a common year through the use of some index. The Department of

---

[3] Cost Considerations in Systems Analysis, Fisher, The Rand Corporation, 1970.

Defense, Bureau of Labor Statistics, and other organizations publish indices for the purpose of adjusting prices to the same year; removing the effects of inflation to see the true trend in real prices; or to project the effects of inflation into the future. Another temporal problem is that companies themselves change over time. Mergers, divestures, business base changes, capital investments[4], accounting system changes, and varying market conditions impact prices, costs, rates, and labor standards. Market conditions vary due to competition, supply, demand, and the life cycle[5] of the product or service. Changes in technology likewise change the cost of goods and services making it problematical to make consistent comparisons except over relatively short time spans. This has the effect of limiting the size of your data set.

The terms and conditions of contracts can also make price comparisons difficult. Did the contract include a warranty, technical support, documentation, training, spares, installation, or delivery costs? What about the contract type? Was it time and materials, cost reimbursable with fixed fee, reimbursable with incentive fee, fixed price with incentive, firm fixed price, or some other arrangement? Was it sole source? Was there a prime contractor, contractor teaming, or did the buying office act as the prime? Did regulatory or environmental considerations impact the contract performance?

The bottom line is this: analysis is more than just crunching numbers; analysis also means understanding the conditions that caused the numbers to be what they are.

Well, isn't that just typical!

Have you ever taken your car in for repair and were told that the repair "typically" costs $200, and wondered, with some concern, exactly what "typically" meant? Words like typically, most likely, usually, commonly, frequently, regularly and on average tend to be used without discretion or precision, which is why you need to determine exactly what the person meant by the term. Was the person referring to the mean or average, which is the sum of the numbers divided by the sample size? Were they describing the median value, which is the middle value of a set of numbers when the numbers have been ranked either from high to low or low to high? Perhaps the person was expressing the value that occurs most often, the mode. Or maybe the answer is none of the above. There is one thing you can know for sure, given any one of these measures the only thing you are really seeing is the tip of the iceberg and you have no idea of what lies beneath the surface. If we are to see the true shape and size of the iceberg we need to understand the amount of dispersion in the data as well as the general shape of the distribution the data comes from.

Think of the dispersion or variation in the data as the size of the iceberg, for sailors, the larger the iceberg the greater the risk. The same is true for analysts working with numbers, the more variation in the data the greater the risk. The range is one way of measuring dispersion.

---

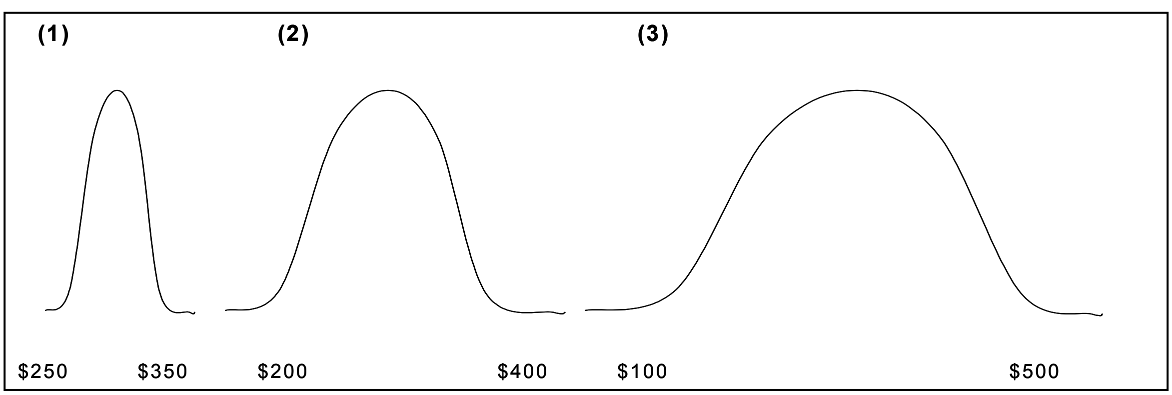
[4] Investments in capital may reduce labor and machine hours while resulting in increases in other direct and indirect costs such as depreciation and maintenance.

[5] The life cycle generally involves a development, growth, maturity, and decline phase. Prices may vary over the life cycle or even within a phase of the life cycle. For example, as a product nears the end of its life cycle the price may fall due to lack of demand or rise due to shortage of supply.

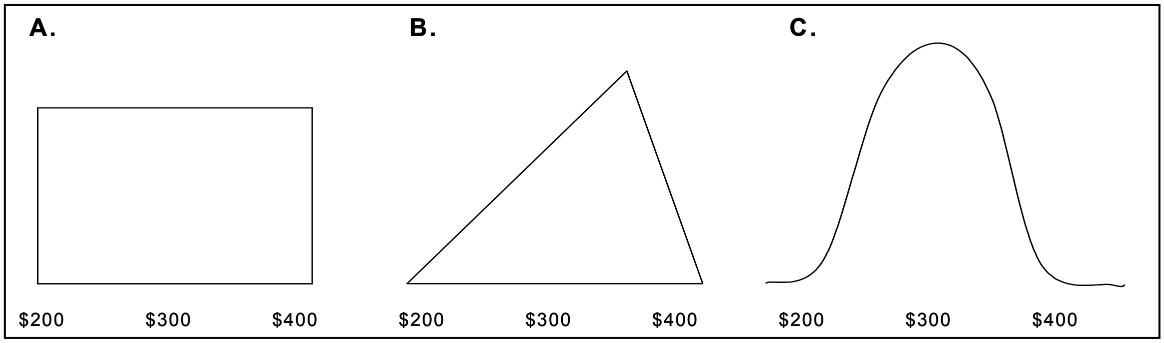
Let’s say that the “typical” car repair of $200 mentioned earlier is the mode, the most frequently paid price. This might actually represent the minimum that you will pay, while at times the repair could run as high as $500. We now know that the range is $300 and that there is a 100 percent probability of the price being $200 or more. However, this still does not tell us if people almost always pay $200 or if there is a fair amount of variation in the prices paid.

The standard deviation is one way of expressing the average or typical variation when the data approximates a normal (bell-shaped) distribution. If the average price of a repair was $300 and the standard deviation $60, then, depending on the sample size, we could assume that about 60 to 70 percent of the prices would fall in the $240 to $360 range. We can use the standard deviation and the T distribution to construct a confidence interval to bound the population mean or with a prediction interval to bound the next observation. If you want to put the standard deviation into perspective, divide the standard deviation by the sample mean to arrive at an average percent variation. The $60 divided by $300 is 20 percent. This calculation is known as the coefficient of variation or CV. We can now make the statement that our average price is $300 and it varies on average by $60 or it varies on average by 20 percent. An alternative measure, which also expresses the average variation in a set of numbers, is the mean absolute deviation or MAD. One advantage of the MAD is that it can be used without making any assumptions of the normality of the distribution.

The mean, median, and mode are all measures of central tendency, and in a distribution that is fairly normal (bell-shaped) they are similar values. However, there are many types of distributions, which is why it is important to not only know whether a value represents the mean, median, or mode, but also to know the distribution that the value came from. The distribution can give you insight to the risk associated with the sample as a whole as well as the risk associated with using any given value from the distribution.



Examples (1), (2), and (3) all have the same mean, median, and mode of $300. However, we can see that the risk in accepting $300 as our position increases as the variability increases as we move from (1) to (2) to (3). We can see here the importance of determining the range of possible values for any given distribution. Asking questions like what are the optimistic, pessimistic, and most likely costs, in addition to asking what factors drive the costs are essential. Of equal importance is the shape of the distribution and where on the distribution your point estimate is located.



In example (A) we see a uniform distribution where any value from $200 to $400 has the same likelihood of occurring. In other words, the price is just as likely to be $225 as it is to be $250 or $350. The mean and median are $300, but there is no mode. If you were quoted a price of $250 for example, there would be a 75 percent probability of the price actually being higher than $250. From the triangular distribution, (B), we would conclude that the most likely price (mode) is around $350, however, there is a high probability that the price would actually be less than $350. The mean and median in this case are probably between $300 and $350. Our normal distribution, (C), has a mean, median, and mode of $300. If the standard deviation were $60, then about 60 to 70 percent of the prices would be between $240 and $360 depending on the sample size. In each case, by understanding the distribution and the assumptions that are associated with the population from which the sample is drawn the analyst can make a statement about the probability or risk associated with any value from the distribution.

I've Got Your Number!

There is a story of three blind men who were trying to describe an elephant. One grasped the elephant's ear and thought surely an elephant must be like a giant bat. The second felt of one of the elephant's legs and thought it must be like a great tree. The third held the elephant's tail and was sure it must be like a snake. Each only had a part of the whole and therefore could not fully describe the elephant. When we ignore differences in definitions, content, format, economic year, market conditions, etc., we are only grasping a part of the data and we will be prone to misunderstand that about which we are trying to make a statement or judgment. Normalization, adjusting the data to make it more consistent, is one of the first steps in data analysis.

The next step is like assembling a jigsaw puzzle; you have to determine what each piece represents and how it will be used. Does the number represent the mean, median, or mode? What is the range in the data? How much variation exists? Do I have all of the pieces? What does the distribution look like and where is a particular value on the distribution?

The final step is to interpret the information you have assembled. Have you ever seen one of those pictures that appears at first to be just a random pattern of lines and textures, then, after staring intently at the pattern for a few moments, you suddenly see a three-dimensional picture within the picture? We must look past the numbers to see the picture within if we are to fully understand the implications of the data and use the data to support management decisions.

Prof Steven Malashevitz
DAU Midwest